通信のクラスタ間遷移に基づくサイバー攻撃検知手法

荒木 翔平 †　　山口 由紀子 ††　　嶋田 創 ††　　高倉 弘喜 ‡

† 名古屋大学 情報科学研究科
464-8601 愛知県名古屋市千種区不老町
araki@net.itc.nagoya-u.ac.jp

†† 名古屋大学 情報基盤センター
464-8601 愛知県名古屋市千種区不老町
{yamaguchi,shimada}@itc.nagoya-u.ac.jp

‡ 国立情報学研究所
101-8430 東京都千代田区一ツ橋 2-1-2
takakura@nii.ac.jp

あらまし サイバー攻撃は年々巧妙化しており，深刻な問題となっている．従来の対策の 1 つであるシグネチャ型侵入検知システムは，あらかじめ登録された攻撃の特徴に一致するものを検知するシステムであるため，未知の攻撃には対応できない．そこで，本稿では攻撃の特徴を必要としないトラフィックの状態遷移に基づくアノマリ型の検知手法を提案する．本手法では，トラフィックデータからセッション単位に特徴量を抽出しクラスタリングを行い，各ホストの通信のクラスタ遷移によってスコアを付け，通信が正常か攻撃かの検知を行う．前後のセッションのクラスタ遷移も含めてスコアリングを行うことで検知精度ををを高めるとともに、閾値により未知攻撃の検知も行う．実験の結果，90%近くの攻撃を検知し，86%の未知攻撃を検知できたことを確認した．

Cyber Attack Detection Based on Cluster Sequence of Communication

Shohei Araki†　　Yukiko Yamaguchi††　　Hajime Shimada††　　Hiroki Takakura‡

†Graduate School of Information Science, Nagoya University
Furo-cho, Chikusa-ku, Nagoya, 464-8601, JAPAN
araki@net.itc.nagoya-u.ac.jp

††Information Technology Center, Nagoya University
Furo-cho, Chikusa-ku, Nagoya, 464-8601, JAPAN
{yamaguchi,shimada}@itc.nagoya-u.ac.jp

‡National Institude of Informatics
2-1-2 Hitotubashi, Chiyoda-ku, Tokyo, 101-8430, JAPAN
takakura@iij.ac.jp

Abstract Cyber attacks have been sophisticating. A signature-based intrusion detection system (IDS) is one of the countermeasures, but it detects only attacks whose signature are provided in advance. In order to detect unknown attacks without signatures, it is expected to develop some method, e.g., an anomaly-based IDS. In this paper, we propose a method that dose not require signatures of attacks. In this method, we perform clustering and give a score by cluster sequence of communication per host after extracting features from communication. We aim to improve the detection accuracy by taking session sequence of prior and latter communication into account.

1 はじめに

インターネットを介したサイバー攻撃の脅威は年を追うごとにより深刻な問題となっている．近年のサイバー攻撃は金銭や情報を不正に入手するために行うことが多くなっており，攻撃を成功させるために攻撃方法は年々巧妙化している．アンチウイルスソフトでは検知できないように細工されたマルウェアを使用したり，正常な通信に似せて外部の Command & Control (C&C) サーバと通信を行ったりするため，一般的なセキュリティソフトでは検知することが困難となってきている．そのため，このような巧妙化したサイバー攻撃に対策を講じる必要性が高まってきている．

従来のサイバー攻撃対策の 1 つに侵入検知システム (IDS:Instrusion Detection System) が存在する．シグネチャ型 IDS は攻撃の通信のシグネチャを登録しておき，そのシグネチャに一致する通信を攻撃の通信として検知するシステムである．低い誤検知率にて高い攻撃の検知率を得ることができるため，商用の IDS でも幅広く利用さてれている．しかし，近年の巧妙化する攻撃に対しては，あらかじめシグネチャを登録することが難しいため，不十分であると言える．一方，アノマリ型 IDS では正常の特徴を学習し，その特徴から外れるものを異常と検知する手法である．正常以外の通信を攻撃として検知するので，ゼロディ攻撃のような未知の攻撃でも検知できる手法であるが，正常な通信を定義することが難しく，シグネチャ型 IDS に比べ誤検知率が高いといった問題が存在する．また，正常であるか，異常であるのかの判断しか行われないため，検知した通信が既知の攻撃ものなのか未知の攻撃なのかの判断がつかないという問題も存在する．日々の攻撃の大半は既知の攻撃であるため，たとえ未知の攻撃を検知していても，膨大なアラートの中から未知の攻撃のものを探すのは困難である．

本稿では，未知の攻撃でも検知できるようなトラフィックの状態遷移に基づくアノマリ型の検知手法を提案する．本手法では，通信の特徴量に対してクラスタリングを行い，ホスト毎のクラスタの遷移による前後の通信も考慮することによって，単一の通信のみを検査している従来手法よりも検知率を高め，誤検知率を低くすることが可能であること示す．また，閾値を設定し未知攻撃の検知も検知できることを示す．アノマリ型 IDS では正常な通信を学習し，その特徴から逸脱する通信を異常として検出するが，本手法では現在の通信に対してクラスタリングを行い，正常か異常かの分類を行う．

2 関連研究

サイバー攻撃の検知に関して，シーケンスを考慮したものとしては，隠れマルコフモデルを用いた手法 [1] や，ルールベースで検知する手法 [2] が存在する．

水谷らは [2] を用い，マルウェア感染端末に対するマルウェアの通信の状態遷移モデルを作成し検知を行う手法を提案した [3]．水谷らはマルウェアの通信傾向の状態遷移モデルを作成することにより，高精度でマルウェアの通信を検知することができることを示した．しかし，多種多様なすべてのマルウェアの通信に対する汎用的なモデルを作成することは難しく，侵入検知システムの構築ににおいて，攻撃モデルの作成は多大なコストがかかり容易ではない．

未知攻撃の検知手法として，Song らはシグネチャ型 IDS から特徴量を抽出し，未知攻撃の検知を行う手法を提案した [4]．ゼロディ攻撃は既存の攻撃コードを改良したり，組み合わせたりして行われることがあり，シグネチャ型 IDS が不自然なアラートを出力することがある．また，既知攻撃に比べて，攻撃の有効性を確認するため長期間攻撃を行ったり，特定のポートのみに攻撃を行ったりする傾向がある．Song らはこれらの未知攻撃の特徴に着目し，シグネチャ型 IDS のアラートから特徴量を抽出し，One-Class SVM を用いて未知攻撃の検知を行った．しかし，この手法では IDS が反応していない未知の攻撃には対応することができない，

したがって，未知攻撃の特徴を反映し，攻撃モデルの作成をする必要のない未知攻撃検知手法が必要である．

3 検知手法

3.1 概要

サイバー攻撃は正常な通信よりも数が少ないと考えられるため，多くの正常な通信のトラフィックの状態遷移とは異なる遷移のものを検知することでサイバー攻撃を検知できると考えられる．また，2 節にて述べたように，未知攻撃には既知攻撃とは違った特徴があり，それらがトラフィックの状態遷移にも反映されると考えられ，攻撃と検知されたものの中でより数の少ないものを未知攻撃として検知できると考える．そこで，本稿ではトラフィックデータにクラスタリングを行い，その中から数の少ない異常なトラフィック遷移を検知する手法を提案する．本手法では，現在の通信に対してクラスタリングを行い，正常の通信と攻撃の通信を分類するため，アノマリ型 IDS のように正常な通信の特徴を学習する必要はない．

提案手法の流れを図 1 に示す．まず，トラフィックデータから特徴量を抽出し，クラスタリングを行う．次にホスト毎にクラスタリングの結果を抽出し，クラスタ列を作成する．そのクラスタ列からクラスタ遷移の頻度を計算し，閾値によって攻撃かどうかの判定を行う，

図 1 内の各ステップは以下の節にて述べる．

- 特徴量抽出 (3.2 節)
- クラスタリング，遷移情報抽出 (3.3 節)
- クラスタ列の分解，スコアリング (3.4 節)

3.2 特徴量の抽出

まず，ネットワークトラフィックデータから特徴量の抽出を行う．抽出する特徴量は京都大学ハニーポットトラフィックデータ Kyoto2006+[1] で使用されているものと同じものを使用する．Kyoto2006+は TCP セッションごとに 14 個の基本特徴と 10 個の追加特徴から構成される．基本特徴は IDS の評価に広く用いられている KDDCup1999[2] の 41 個の特徴量の中で特に重要な特徴である 14 個を使用している．基本特徴では接続時間，送受信バイト数，サービスタイプ，接続状態，接続回数などを特徴量として抽出している．本研究では Kyoto2006+の基本特徴の中で，数値属性ではない「サービスタイプ」と「接続状態」を除いた 12 個の特徴量を使用する．

特徴量の尺度は特徴量によって異なるため，抽出した特徴量は平均 0，分散 1 のデータに正規化する．全セッションデータ数を N，セッションデータ $x_i = \{x_{i1}, x_{i2}, \ldots, x_{id}\}(1 \leq i \leq N)$ とし，まず，特徴量毎に平均 μ_j と分散 σ_j を求める．

$$\mu_j = \frac{1}{N}\sum_{i=1}^{N} x_{ij}$$

$$\sigma_j = \frac{1}{N-1}\sum_{i=1}^{N}\sqrt{(x_{ij}-\mu_j^2)}$$

その後，$\boldsymbol{x_i}$ の各特徴量に対して，正規化後の値 $\boldsymbol{x_i'}$ を求める．

$$x'_{ij} = \frac{x_{ij}-\mu_j}{\sigma_j}$$

3.3 クラスタリング

抽出した特徴量を元にクラスタリングを行う．代表的なクラスタリング手法には，k-means 法が存在するが，k-means 法ではパラメータにクラスタ数を設定しなけれ ばならなかったり，各クラスタが超球状で同じくらいの半径であることが暗黙的に仮定されているため，多種多様なデータが存在し，あらかじめクラスタ数を定義することができないセッションデータには適していないと考えられる．

提案手法ではクラスタリングに Density-Based Spatial Clustering of Applications with Noise (DBSCAN)[5] を用いる．DBSCAN ではデータの密度を基準にクラスタリングを行い，あらかじめクラスタ数を指定する必要がない．また，どのクラスタにも属さないデータをノイズとして判定することができるため，セッションデータに適していると考えられる．

DBSCAN は距離 eps と最小ポイント数 $MinPts$ の 2 つのパラメータを取る．ある点 p から距

[1] http://www.takakura.com/kyoto_data/

[2] http://kdd.ics.uci.edu/database/kddcup99/dkkcup99.html

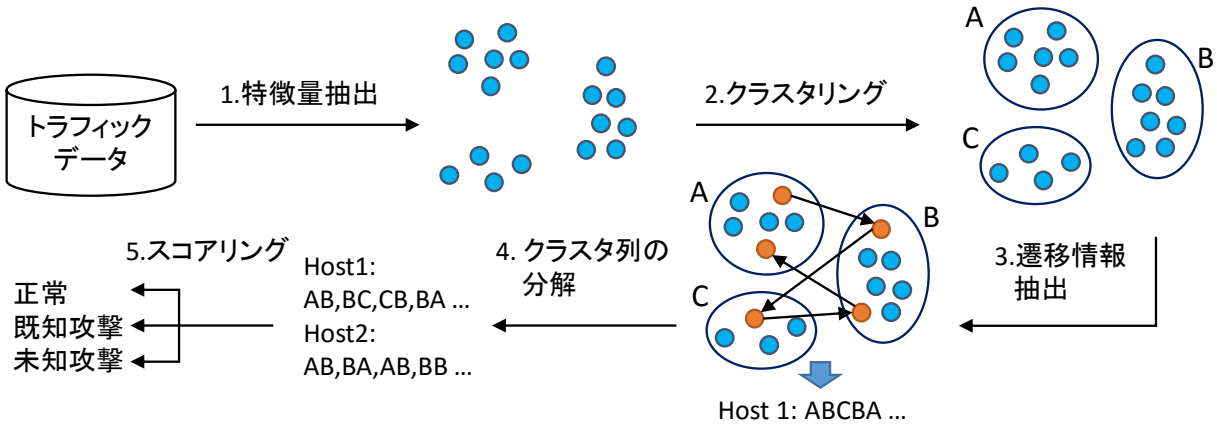

図 1: 提案手法の流れ

離 eps 内にある点集合を近傍 $N_{eps}(p)$ と定義し，p から以下の条件を満たす到達可能点 q を見つける．

- $q \in N_{eps}(p)$
- $|N_{eps}(p)| \geq MinPts$

ある点から，この到達可能な関係を推移的にたどっていくことによって得られる点集合を 1 つのクラスタとし抽出する．また，すべての点から到達できなかったり，近傍に $MinPts$ 未満の数の点しか存在しない点をノイズデータとして扱う．ノイズデータは多くのデータが所属するクラスタのデータではなく，外れ値に所属するデータと考え，本手法ではすべてのノイズデータが「外れ値」という同一のクラスタに属するとする．

全セッションデータについてクラスタリングを行った後，クラスタリング結果をホスト毎に抽出する．ある IP アドレスがセッションデータの送信元もしくは宛先 IP アドレスに含まれていた場合，そのセッションデータはその IP アドレスのホストが行った通信とみなす．この操作によって，ホスト毎にクラスタ列を作成する．また，通信の前後関係を崩さないようにクラスタ列はホストの通信を時系列順に並べる．

3.4 スコアリング

すべてのホストのクラスタ列を bi-gram 法にて分解し，クラスタ遷移を抽出する．ホスト h のクラスタ列を bi-gram にて分解した遷移の i 番目のものを b_{hi} とする．また，クラスタ列の i 番目のものを c_{hi} とする．例えば．ホスト h が ABCB. . . と遷移する時，$b_{h2} = \mathrm{BC}, c_{h2} = \mathrm{B}$ となる．

このクラスタ遷移について，全ホストのクラスタ列の中での出現割合をそれぞれ計算し，$p(k)$ とする．例えば，クラスタを AB と遷移する割合は $p(\mathrm{AB})$ と表す．また，クラスタ A からクラスタ B に遷移する割合を $q(B|A)$ とする．

ホスト h の i 番目のセッションのスコア S_{hi} を以下の式にて計算する．このスコアは高いほど正常な通信であるとみなす．

$$S_{hi} = \sum_{k=i-l+1}^{i+l} p(b_{hk}) q(c_{hk+1}|c_{hk})$$

l はスコアリングにおいて用いるセッションの数を表し，スコアに影響を与えるセッションの数を制限する．スコアリングにおいて，セッションデータ内に多く含まれているクラスタ遷移は正常な通信のものとして考える．$p(b_{hk})$ はクラスタ遷移の出現割合であるため，多く存在するクラスタ遷移のスコアの値は高くなる．$q(c_{hk+1}|c_{hk})$ はあるクラスタから別のクラスタへ遷移するもののうち，それがクラスタ内でどれだけの割合であるのかを示している．そのため，あるクラスタ内から同じような遷移するパターンが多く存在する場合，スコアが高くなる．

セッションには送信元と宛先ホスト存在するため，セッションのスコアは送信元ホストのス

コアと宛先ホストの平均とする．もし，送信元か宛先ホストのいずれかのホストの通信回数がセッション数 l 未満だった場合，スコアが計算できる l 以上のセッション数を持つホストのスコアをそのセッションのスコアとする．両方のホストとも l 未満だった場合は本手法では判定することができないため，正常の通信と判定する．

スコアが低いものは，セッションデータ内において数の少ない遷移パターンであると考えれられるので，攻撃のセッションとみなす．攻撃の検知には閾値 α を設定し，α 以下のスコアの通信を攻撃として検知する．また，値が α より小さい閾値 β を設定し，スコアが β 以下となる通信について未知の攻撃として検知する．

4 評価実験

4.1 概要

実環境にて正常なセッションと攻撃のセッションを正確に分類し，ラベル付けしたデータを収集するのは困難であるので，Kyoto2006+を用いて評価実験を行う．

Kyoto2006+はハニーポットネットワークで収集されたセッションデータであるため，実環境に比べて攻撃のデータの割合が非常に多くなっている．そのため，攻撃のセッションデータ数が正常なセッションデータ数に対して約 1%の割合となるように抽出し実験を行う．攻撃を抽出する際にはホスト毎のセッションの時間関係をできるだけ崩さないようにするため，ホスト毎にセッションデータを抽出し，攻撃の割合を調整する．また，提案手法にて判定することができない送信元ホストと宛先ホストの両方が l 未満のセッション数しか存在しないセッションデータについてはあらかじめ取り除く．

本実験では，Kyoto2006+の 2009 年 8 月 10 日，20 日，30 日のデータを使用する．正常なセッションの数はそれぞれ 57734，59024，51639 であり，攻撃のセッションの数は前述したように調整し，それぞれ 578，600，523 とした．

4.2 結果・考察

3 つのデータに関しての検知結果を表 2 に示す．パラメータは距離 eps，最小ポイント数 $MinPts$，閾値 α，セッション数 l をそれぞれ 0.5，3，0.05，5 とした．なお，検知結果と実際の分類の関係を表 1 のように定義した時，検知率 (DR) と誤検知率 (FPR) を以下のように定義する．

表 1: 検知結果の定義

		実際の分類	
		攻撃	正常
検知結果	攻撃	TP	FP
	正常	FN	TN

$$\text{検知率 } (DR) = \frac{TP}{TP + FN}$$

$$\text{誤検知率 } (FPR) = \frac{FP}{TN + FP}$$

表 2 より，日付によって多少のばらつきはあるものの，概ね低い誤検知率にて高い検知率を達成できていると言える．

表 2: 攻撃の検知結果 (%)($eps = 0.5, \alpha = 0.05$)

	2009/08/10	2009/08/20	2009/08/30
DR	77.68	79.83	94.07
FPR	7.14	4.67	3.39

図 2 に閾値 $\alpha = \{0.001, 0.005, 0.01, 0.05, 0.1\}$ と変化させた場合の ROC 曲線を示す．また，従来手法との比較として One-Class-SVM (OC-SVM) を用いて攻撃の検知を行った場合の ROC 曲線も示す．OC-SVM は 2009 年 8 月 1 日のセッションデータを学習データとして使用した．実線が提案手法の ROC 曲線，点線が OC-SVM による ROC を表している．

図 2 より，誤検知率が高くなってしまうものの，どの日付でも 90%近い検知率を達成できていることがわかる．特に高い 8 月 30 日に関しては，1%未満の誤検知率にて 90%以上の検知率を得ることができ，非常に高い精度にて検知

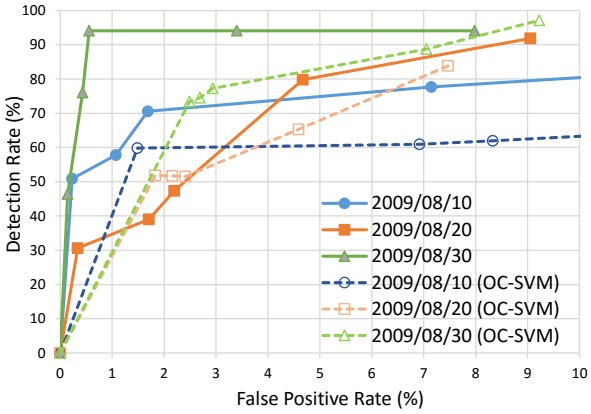

図 2: 提案手法と OC-SVM の ROC 曲線

できていると言える．また，どの日付に対しても OC-SVM による検知と比較して，提案手法の方が低い誤検知率にて高い検知率を達成することができている．

閾値 α について，どの日付でも α の値を小さくすると検知率は低くなってしまうものの，非常に低い誤検知率にて 50%前後の検知率を得ることができる．ネットワーク管理者がパラメータを調整することで，検知率を重視してアラートを出力するか，誤検知率を低く抑えて本当に攻撃のアラートのみを出力するようにするかを選択することができると言える．

表 3 に 3 つのデータの距離 *eps* を変化させた場合の検知率と誤検知率の変化を示す．また，表 4 にそれぞれの *eps* の場合のクラスタの数を示す．*eps* はあるクラスタからユークリッド距離にて *eps* までの近傍点を同一のクラスタとみなすものである．*eps* を大きい値にすることで，特徴量空間でより離れた位置に存在していても，同一クラスタに属するとみなされるようになる．そのため，*eps* が小さい値の時では，正常なセッションが属しているクラスタと攻撃のセッションが属しているクラスタがうまく分離できていたものが，*eps* を高くしたことによって同一のクラスタに属してしまうことがある．8 月 20 日はこのために検知率に悪影響を及ぼしてしまったと考えられる．

一方，*eps* を小さくしすぎるとクラスタ数が増えすぎたり，ノイズデータのクラスタに属するセッションデータが増えたりしてしまう．*eps* が 0.1 の時，クラスタ数が膨大な数になってしまい，クラスタ遷移のパターンが激増してしまい，うまく検知することができなくなってしまった．これらの結果より，検知率と誤検知率は閾値 α でも調整できるため，攻撃を見逃さずに検知するためにも *eps* は 0.5 が適していると言える．

表 3: *eps* による検知率の変化 (%)

	2009/08/10		2009/08/20		2009/08/30	
eps	DR	FPR	DR	FPR	DR	FPR
1.0	70.07	0.26	33.50	0.18	94.07	0.33
0.75	77.68	6.92	34.50	1.37	94.07	0.35
0.5	77.68	7.15	79.83	4.67	94.07	3.40
0.25	94.64	15.15	89.33	7.81	94.07	8.12
0.1	94.64	21.80	81.50	96.08	18.16	96.25

表 4: *eps* によるクラスタ数の変化

eps	2009/08/10	2009/08/20	2009/08/30
1.0	30	31	31
0.75	45	37	44
0.5	52	51	46
0.25	86	78	92
0.1	229	1059	836

4.3 未知攻撃に対する検知結果

2008 年 8 月 30 日のセッションデータについて未知攻撃に対する検知結果を表 5 に示す．攻撃検知の閾値 α，β はそれぞれ 0.05，0.005 とした．距離 *eps*，最小ポイント数 $MinPts$，セッション数 l に関しては 4.2 節の実験同様それぞれ 0.5，3，5 とした．また，表 5 における未知攻撃の検知率 (U_DR) と未知攻撃の誤検知率 (U_FPR) は以下のように定義する．

$$U\_DR = \frac{\text{正しく未知攻撃と検知できた数}}{\text{未知攻撃の総数}}$$

$$U\_FPR = \frac{\text{誤って未知攻撃と検知した数}}{\text{未知攻撃以外の総数}}$$

表 5 より，2%未満の低い誤検知率にて 80%以上の高い検知率にて未知の攻撃を検知することができた．誤検知率は低いものの，検知した未

表 5: 未知攻撃に対する検知結果

		実際の分類		
		未知	既知	正常
検知結果	未知	13	385	223
	既知	2	92	1532
	正常	0	31	49884
U_DR		86.67%		
U_FPR		1.17%		

知攻撃の数に比べ誤検知したものの数は多く，その中から真に未知攻撃のものを探すのは困難ではあるが，ネットワーク管理者がすべてのトラフィックを解析して未知攻撃を調査する手間に比べれば，未知攻撃の可能性のあるトラフィックを大幅に絞り込むことができたと言える．

5 おわりに

本稿では，セッションデータをのクラスタ遷移に着目して，その中から数の少ないパターンを検知する手法を提案した．提案手法では，密度ベースのクラスタリングである DBSCAN を用いて，セッションデータを適切にクラスタリングし，数の少ないパターンに低いスコアを設定し，攻撃として検知した．攻撃のモデルの作成や正常の通信の定義をする必要がないため，幅広い攻撃を検知するとこができ，様々なネットワークに対して適用することができると考えられる．

京都大学ハニーポットトラフィックデータ Kyoto2006+を用いた評価実験では，サイバー攻撃に対して低い誤検知率にて高い検知率を得ることができた．また，未知攻撃の検知に関しても，2%以下の検知率にて 80%以上の検知率を得ることができた．

本提案手法では全データに対しての遷移の割合にてスコアを設定したため，正常なデータのみの環境では誤検知が高くなってしまう可能性がある．また，クラスタリングの結果にて攻撃の検知を行うため，トラフィックデータが集まらなければうまく検知することができない．本手法はアノマリ型 IDS のような正常の特徴の学習が必要ないことは 1 つの利点ではあるが，今後の課題としては，インシデントに対して迅速に対応するためにも，クラスタリング手法を応用してセッションの状態遷移を抽出･学習し，リアルタイムに検知を行う手法を開発することが挙げられる．

謝辞

本研究は平成 25 年度総務省情報通信技術の研究開発「サイバー攻撃の解析・検知に関する研究開発」の支援を受けている．

参考文献

[1] Khanna, R. and Liu, H.: Control theoretic approach to intrusion detection using a distributed hidden Markov model, *Wireless Communications, IEEE*, Vol. 15, No. 4, pp. 24–33 (2008).

[2] Mizutani, M., Shirahata, S., Minami, M. and Murai, J.: ROOK: Multi-session Based Network Security Event Detector, *Applications and the Internet, 2008. SAINT 2008. International Symposium on*, pp. 48–54 (2008).

[3] 水谷正慶，武田圭史，村井　純：Web 感染型悪性プログラムの分析と検知手法の提案，電子情報通信学会論文誌 B， Vol. 92, No. 10, pp. 1631–1642 (2009).

[4] Song, J., Takakura, H. and Kwon, Y.: A Generalized Feature Extraction Scheme to Detect 0-Day Attacks via IDS Alerts, *The 2008 International Symposium on Applications and the Internet(SAINT2008)*, pp. 55–61 (2008).

[5] Ester, M., Kriegel, H.-P., Sander, J. and Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise., *Kdd*, Vol. 96, No. 34, pp. 226–231 (1996).